DOCUMENT RESUME

ED 223 720 TM 820 868

AUTHOR Moy, Raymond
TITLE Score Equating and Nominally Parallel Language Tests.
PUB DATE May 82
NOTE 35p.; Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (16th, Honolulu, HI, May 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; Language Proficiency; *Language Tests; Latent Trait Theory; Scaling; Statistical Analysis; *Test Reliability; *Test Theory
IDENTIFIERS Equipercentile Equating; Generalizability Theory; Linear Equating Method; Parallel Forms Reliability; *Parallel Test Forms

ABSTRACT

Score equating requires that the forms to be equated are functionally parallel. That is, the two test forms should rank order examinees in a similar fashion. In language proficiency testing situations, this assumption is often put into doubt because of the numerous tests that have been proposed as measures of language proficiency and the heterogeneous nature of the examinee population to whom the tests are given. A methodology is proposed for handling the problem of equating nominally parallel language tests. In particular, two multiple-choice grammar tests -- one whose items are based on academic texts and cover a wide range of grammar topics, and the other whose non-academic vocabulary and contexts focus on a narrower range of grammar topics -- are examined for functional parallelism through a generalizability study. It was found that these surface differences did not cause any significant reordering of examinee ranks and that score equating could take place. Three commonly used score equating methods, namely those based on linear, equipercentile, and latent trait models, were then applied and evaluated for appropriateness to this particular equating situation. The equipercentile method was found to provide the most accurate and stable results. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223720

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Score Equating and Nominally Parallel Language Tests

by

Raymond Moy

University of California, Los Angeles

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Moy

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

11-82-59 TM 820 808

A paper presented at the 16th annual meeting of TESOL, Honolulu, Hawaii, May 1982.

Abstract

Score equating requires that the forms to be equated are functionally parallel. That is, the two test forms should rank order examinees in a similar fashion. In language proficiency testing situations, this assumption is often put into doubt because of the wide range of tests that have been proposed as measures of language proficiency and the heterogeneous nature of the examinee population to whom the tests are given.

In this paper, a methodology is proposed for handling the problem of equating nominally parallel language tests. In particular, two multiple-choice grammar tests -- one whose items were based on academic texts and covering a wide range of grammar topics, and the other using non-academic vocabulary and contexts focusing on a narrower range of grammar topics -- are examined for functional parallelism through a generalizability study.

It was found that these surface differences did not cause any significant reordering of examinee ranks and that score equating could take place. Three commonly used score equating methods, namely those based on linear, equipercentile, and latent trait models, were then applied and evaluated for appropriateness to this particular equating situation. The equipercentile method was found to provide the most accurate and stable results.

Introduction

When one attempts to equate the scores from two different test forms, one is making an assumption that the forms are measures of the same thing (See Angoff, 1971, for a more complete explanation). This is problematic for many language tests because it is unknown how similar the tests need to be before they can be considered to be functionally parallel. For example, one would like to claim that the Michigan Test of English Language Proficiency is measuring the same thing as the TOEFL exam, but when one considers format differences and specific item content differences, claims to parallelism are suspect. The equating problem, then, actually consists of two parts. First of all, one must demonstrate the functional parallelism of the two test forms, and only then can one derive a system of equated scores.

In this paper, it will be demonstrated how two grammar subtests, which came from separate item pools, were tested for parallelism via generalizability theory, and then how a system of appropriate equated scores was derived. The methodology to carry these steps out are applicable to any equating problem. In particular, generalizability theory forces the test administrator to consider how his tests systematically differ from one another and whether these sources of differences are relevant to test performance.

The Measurement Instruments to Be Equated

For the Fall 1981 UCLA English as a Second Language Placement Exam (the ESLPE), two grammar subtests were administered. One of the subtests consisted of old grammar items held over from the previous

school year while the other subtest contained newly designed items.

The old grammar items were developed to match the grammar objectives of the ESL service courses at UCLA. All items were of the multiple-choice format with full sentence contexts. These items were independently critiqued by a panel of ESL specialists for plausibility of distractors, difficulty level, and ambiguities. What resulted was a set of items that, from a face validity point of view, were appropriate for placement purposes.

In Figure 1 are the resulting score distribution, mean, standard deviation and test reliability based on 278 examinees tested in the Spring of 1981. The distribution of scores is quite typical of those obtained from the grammar subtests of previous forms of the ESLPE.

Insert Figure 1 about Here

For the Fall 1981 test, a subsample of 30 of these items was used, but, in addition, 40 new items were introduced. These new items differed from the old items in two significant respects. First of all, the item stems were more academic in nature, having been based on sentences taken from college level texts. Secondly, no control was placed on the grammar areas being tested. However, the multiple-choice format was retained and the same process of item review and revision was followed to maintain face validity.

To get a clearer idea of the content differences of the two grammar

test forms, the items were classified according to what the correct answers were and under which section of Quirk & Greenbaum's Concise Grammar of Contemporary English (1973) the correct answer for an item would best be explained. The results are shown in Table 1. Across the top are the chapter numbers from Quirk & Greenbaum and under each are the item numbers and the chapter subsections in which the item content is treated. For example, under Chapter 3, on the verb phrase, is item 31 which is treated under subsection 3.28.

Insert Table 1 about Here

As can be seen, there seems to be a greater variety of items in the new test form, but there is a similarity between the forms in that both contained relatively more items from Chapters 3 4, and 11. These were the chapters on the verb phrase, the noun phrase, and the complex sentence. Figures 2 and 3 contain the respective distributions of the old and new grammar tests for the Fall 1981 test. Each distribution is based on the scores of 629 examinees. Of note, there is a slight ceiling effect for the old grammar items and the distribution is more negatively skewed than the new grammar distribution. However, if handled properly, both problems are treatable by one or more score equating procedures.

Insert Figures 2 & 3 about Here

Given the content differences between the forms, the academic origin of the new items, and the score distribution differences, there is cause for doubt that the two tests are functionally parallel and that score equating can take place. Therefore, it is first necessary to determine the extent to which the forms are testing the same ability. Because all examinees took both subtests it was possible to examine how closely the two tests appear to measure the same ability through a generalizability study.

The Generalizability Study

In a generalizability study, examinees are observed under representative measurements obtained across conditions that the researcher suspects may contribute to score variance (For a review of generalizability theory, see Shavelson & Webb, 1981). For instance, if the ability being tested by items of the first grammar test is different from those of the second test, then one would expect score variation due to test form. On the other hand, if the abilities are the same, the test form distinction should contribute little or nothing toward explaining score variation. Observed score variance over persons, then, can be broken down into variance components and estimation of these components are obtained from the formulas for the expected mean squares of an analysis of variance. The design of the ANOVA used in this study is shown in Figure 4. Here, there are two measurement facets of interest: test form and grammar area subtests. The questions to be answered are whether it is necessary for the grammar item developers to make a distinction between those forms based on college texts and those which are not, and whether it is important

that test items be designed with particular grammar objectives kept in mind. The null hypotheses being tested are that the item pools, as represented by the two test forms, are interchangeable with one another and the grammar subtests within and between each form are also interchangeable.

Insert Figure 4 about Here

One other concern that language test constructors often have that is related to the interchangeability of tests hypothesis is that there should be no bias with respect to the examinee's language learning background. For example, a group of examinees may benefit or suffer from the differences in the tests. If the tests are parallel, no such advantage should be observed. To examine this problem, subjects were classified into three groups according to their language learning background and this factor was introduced into the generalizability study. The first group consisted of European language university students, the second group were the non-European university students, while group 3 were all non-academic examinees, i.e., those whose goal is simply to learn English and not to obtain a university degree. If the tests are measures of the same construct, there should be no group by test form interactions.

Generalizability studies are based on the principle that score variations for persons can be decomposed into identificable sources. Working under the assumption that all test items are interchangeable in sampling

examinee test ability, any differences in test performance across these items is considered to be due to error of measurement. This error is further decomposed into systematic error due to measurement conditions deemed irrelevant to the ability to be measured and random error due to chance fluctuations. Examples of systematic error sources might be something like date of measurement, testing site, test administrator, or test form. Since we know that all of these factors may affect test performance, we need to obtain estimates of the extent to which they attribute to overall score variation. If the systematic error variation is large, then the test administrator must sample the examinee's behavior over several conditions of the error sources before drawing any conclusions about the examinee's true ability. On the other hand, if the variation is small, then ability measured under one condition is seen as similar to that of ability measured under another.

In order to obtain variance component estimates for the design in Figure 4, it was necessary to run a balanced designed ANOVA. That is, equal cell-sizes and equal numbers of items within each subtest. To do this, 135 examinees from each group were randomly selected and five-item subtests were formed from the verb phrase, noun phrase, and complex sentence grammar areas. Because many subtest areas did not contain five items, it was not possible to include all the subtest areas in the generalizability study. However, the variance component for the three subtests which were included could be interpreted to be representative of the variation present in the overall subtest universe.

Table 2 contains the ANOVA results, and these results generally support the interchangeability hypotheses. Along the left-hand side are the sources

of score variation. G stands for group, F for form, S for subtest, P(G) for persons within groups, GF for the interaction of groups by form, etc.

Insert Table 2 about Here

No significant effects for form or subtest were observed. Nor was there any interaction between groups and test form or groups and subtests, implying that the test forms are functionally parallel across groups.

Table 3 contains the estimates of the variance components. Because the scale of measurement is at the item level, the variance is limited to between 0 and 1. The total observed variance at the item level is .218, thus the relatively small values for the variance components. However, what is of interest is the relative sizes of the components rather than their absolute sizes. The largest components are observed for group, person, item, and residual, which is the PI(GFS) term.

Insert Table 3 about Here

One can use these components to predict the reliability of a test N items long. As the number of items, subtests, and forms goes up, the variance of mean performance across these facets of measurement goes down and test reliability should go up. If the hypothesis of subtest

or form interchangeability is true, one should be able to reproduce the observed reliability coefficient of an N item test from the variance component estimates -- that is, one should be able to reproduce the reliability coefficients of the actual observed old and new grammar tests from the components of variance estimated from the generalizability study on the limited samples of persons, subtests, and items.

Reliability is defined as the proportion of true score variance to observed score variance. This is illustrated as follows for a two form, three subtest, 30 item test:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pf}^2}{n_f} + \frac{\sigma_{ps}^2}{n_s} + \frac{\sigma_{pfs}^2}{n_f n_s} + \frac{\sigma_{pi(fs)}^2}{n_i n_f n_s}}$$

$$= \frac{.017}{.017 + \frac{.000}{2} + \frac{.001}{3} + \frac{.001}{2*3} + \frac{.144}{5*2*3}}$$

$$= .76$$

The variance component for persons is the true score variation which goes in the numerator, and in the denominator is observed score variance which is composed of true score variation plus error variation, which in this case is any interaction involving persons. Each of these interaction components is divided by the number of conditions of measurement. The resulting reliability coefficient is .76. The discrepancy between this

coefficient and that of .824 of the 30 item old grammar test in Figure 2 is due to the fact that this higher coefficient includes the group variance as true variance. If this is included in the reliability computations along with any interaction components involving groups going into error, then the reliabiltiy coefficient (based on $n_f = 1$, $n_s = 3$, and $n_i = 10$) is estimated to be .823 which is nearly the same as that of the old grammar test. For a forty item test the computed reliabiltiy would be .851 which again approximates the reliabiltiy of the new grammar test which was .872.

The implication of all of this is that it apparently does not matter which subtest or test form is used. That is, the academic item stems seemed to be of equal effectiveness as the non-academic items in discriminating between student abilities. Also, the subtest classifications do not seem to play a significant part in determining the difficulty of a test.

If one can eliminate test form differences as a source of discrepancy between test performance on two different test forms, then score equating can take place since the only variation source would presumably be random error measurement unreliability.

Selection of a Score Equating Procedure

The next step in the equating process, then, is to select an appropriate equating method. In general, three models are commonly used, but each model has its own set of assumptions in addition to the principal assumption of test form functional parallelism.

Perhaps the simplest method of equating is linear equating in which standard scores from each form are set equal to each other. This is represented by Equation 1 where both X and Y scores are adjusted for

differences in means and standard deviations. The underlying assumption behind linear equating is that the distribution of scores on the two forms are identical.

$$1) \quad \frac{X - \bar{X}}{s_x} = \frac{Y - \bar{Y}}{s_y}$$

The second method of equating does not assume equal distributions, but that persons of equal ability will have equal percentile scores. For example, if a person scores in the top 10 percent of one distribution, he or she should score in the top 10 percent of the other distribution. This method of equating assumes that the populations upon which the percentile scores are based are the same.

The third method of equating is based on latent trait modeling and does not depend on either equal score distributions or equivalent standardizing populations (see Lord, 1980, or Warm, 1978, for more thorough explanations of latent trait modeling). But it does make a very strong assumption that the items in the test forms are unidimensional in nature. That is, a person's performance on the items is determined by a single underlying ability. It is also assumed that the probability of a person answering the items correctly is described by curves similar to those in Figure 5. The horizontal axis is the ability scale while the vertical axis is the probability of getting an item correct.

Insert Figure 5 about Here

These curves are normal ogive curves and they show that as a person's ability increases, the probability of getting the items correct also increases and approaches 1. The general equation to describe curves like these is Equation 2.

2) $p_i(\theta_j) = c_i + (1 - c_i)/(1 + \exp(-1.7a_i(\theta_j - b_i))$

where:

$p_i(\theta_j)$ = probability of person j with ability θ getting item i correct.

a_i = discriminating power of item i.

b_i = difficulty level of item i.

c_i = lower asymptote of the item response function.

θ_j = ability level of person j.

Although the formula looks complicated, the way it works is quite straightforward. Changing b_i, item difficulty, simply moves the whole curve horizontally along the ability scale. For example, Item 2 is a more difficult item than Item 1, and it is further to the right. Changing a_i, item discrimination, changes the steepness of the curve. Item 2 is steeper than Item 1. Finally, changing c_i, sometimes referred to as the guessing parameter, determines the lower asymptote of the curve. For example, if a person with very low ability were presented with a four-alternative multiple-choice item, his probability of getting it correct by chance would be one out of four. The lower asymptote should be .25. For Item 2 the lower asymptote approaches 0 while for Item 1,

it approaches about .4.

In sum, for latent trait equating, once the item curves have been established, it does not matter which items are used to estimate ability because all items have already been linked to the same ability scale. The particular latent trait model used in this study was the Rasch model in which the lower asymptote of all items is assumed to be 0, and all items have equal discriminability.

Evaluation of the Score Equatings

Because of the differences in score distributions of the two grammar tests, it appears that the equipercentile or latent trait methods would be best. Figure 6 shows the score equatings obtained by the three different methods. On the horizontal axis are scores on the new grammar test, while on the vertical axis are the scores on the old grammar test. Each equating model generates a different line to relate old scores to new scores. For example, using the linear equating line, a score of 10 on the old test corresponds to a score of about 5 on the new test, while by equipercentile equating, 10 corresponds to about 9 on the new test.

Insert Figure 6 about Here

As an index of equating appropriateness, the average difference between an equated score and an actual score can be used. Although the equatings look quite different, the average differences between actual scores and equated scores is pretty much the same across methods.

For the linear equating, this average difference is 4.96 while for the equipercentile equating it is 4.68, and for the Rasch equating, it is 4.87. The similarity of the index figures is probably due to the fact that the equatings are similar in the upper ranges, and that relatively fewer cases fall in the low score ranges where the most discrepancy among the equating methods takes place. However, the linear equating seems to be the least appropriate equating model.

If one examines the spread of difference scores around the equating line, the inadequacy of the linear equating becomes more apparent. Figure 7 shows the scatterplot of equated minus actual scores around the equating line. The difference scores should cluster around the 0-line axis, which is the expected value of an equated score minus an actual score.

Insert Figure 7 about Here

As can be seen the linear equating grossly underestimates an actual new grammar test score from an old grammar score in the lower ranges. While in Figures 8 and 9, containing similar scatterplots for the equipercentile and Rasch equatings, this effect is reduced. These figures show, however, that there still may be problems in the equating of scores from the lower ranges of ability. Often associated with low test scores is the increased effect of guessing and other aberrant test behaviors that will affect scores in unpredictable ways. It may be possible to

improve on the Rasch equating by allowing the lower asymptote parameter to vary, instead of being set to 0.

Insert Figures 8 & 9 about Here

Despite the problems at the lower ranges, the upper range equatings seem to be quite adequate. In Figure 10 are equipercentile equating lines derived separately from the European, non-European, and non-academic examinee groups. All these equipercentile lines are similar to one another, but they appear to be most stable for the higher scores. Stability of equatings across examinee groups is the main characteristic one should look for as an indication of test form functional parallelism as well as equating model appropriateness.

Conclusion

When the test administrator is faced with equating the performance of students across two tests, it is necessary that he or she be convinced that the tests are functionally parallel. The grosser the surface differences, the more one is obligated to carry out a systematic study of those differences such as with a generalizability study. Although generalizability studies require planning and a working knowledge of applied statistics, the effort is well worth it for understanding what your tests are measuring and what you can do with them in terms of decision making and score equating.

References

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum, 1980.

Quirk, R. & Greenbaum, S. A concise grammar of contempory English. New York: Harcourt Brace Jovanovich, 1973.

Shavelson, R. J. & Webb, N. M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 1981, 34, 133-166.

Warm, T. A. A primer of item response theory. (USCG Tech. Rep. No. 941278) Springfield, Va.: National Technical Information Service, U. S. Department of Commerce, 1978.

Table 1

Distribution of Fall 1981 Grammar Items by Content Area
Classified by Quirk & Greenbaum Chapters

	3	4	5	6	7	8	9	10	11	12	13	14
Old Item Pool (30 items)	31-3.28 50-3.28 32-3.29 34-3.30 44-3.30 48-3.32 33-3.33 38-3.37 37-3.48	45-4.13 47-4.13 60-4.13 46-4.19 42-4.88	49-5.8	57-6.7 55-6.8			59-9.16		52-11.7 40-11.14 54-11.18 36-11.25 56-11.31 39-11.48	53-12.3 58-12.4 51-12.19 41-12.21	43-13.11	35-14.18
New Item Pool (40 items)	73-3.22 88-3.24 78-3.28 80-3.28 74-3.30 82-3.30 69-3.48 70-3.50 71-3.51	64-4.13 65-4.13 61-4.20 62-4.20 85-4.89 77-4.90 63-4.92	96-5.24 66-5.27 67-5.37	72-6.7 86-6.13 97-6.21 98-6.27	87-7.5 89-7.39	94-8.20	100-9.20	95-10.13 93-10.24 91-10.43	68-11.7 75-11.13 99-11.16 92-11.18 83-11.30 79-11.37	76-12.27	90-13.8 84-13.11 81-13.14	

Table 2

ANOVA Table for the Generalizibility Study

Source	SS	df	MS	F	error term
G	105.05	2	52.52	35.01*	GF+P(G)-PF(G),df=5
F	40.33	1	40.33	2.06	I(FS)+PF(G)-PI(GFS), df=24
S	20.29	2	10.15	8.75	FS+PS(G)-PFS(G), df=2
P(G)	268.16	402	.67	4.53*	PF(G)
GF	1.97	2	.98	1.63	PF(G)+GI(FS)-PI(GFS), df=49
GS	8.27	4	2.07	2.61	GFS+PFS(G)-PI(GFS), df=4
FS	2.30	2	1.15	.06	I(FS)+PFS(G)-PI(GFS), df=24
I(FS)	470.48	24	19.60	135.94*	PI(GFS)
PF(G)	59.24	402	.15	1.02	PI(GFS)
PS(G)	127.43	804	.16	1.04	PFS(G)
GFS	3.11	4	.78	1.30	PFS(G)+GI(FS)-PI(GFS), df=49
GI(FS)	28.39	48	.59	4.10*	PI(GFS)
PFS(G)	122.43	804	.15	1.05	PI(GFS)
PI(GFS)	1391.31	9648	.14		

*p < .05

Table 3

Estimates of Variance Components

G	.013
F	.003
S	.002
P(G)	.017
GF	.000
GS	.001
FS	.000
I(FS)	.048
PF(G)	.000
PS(G)	.001
GFS	.000
GI(FS)	.003
PFS(G)	.001
PI(GFS)	.144

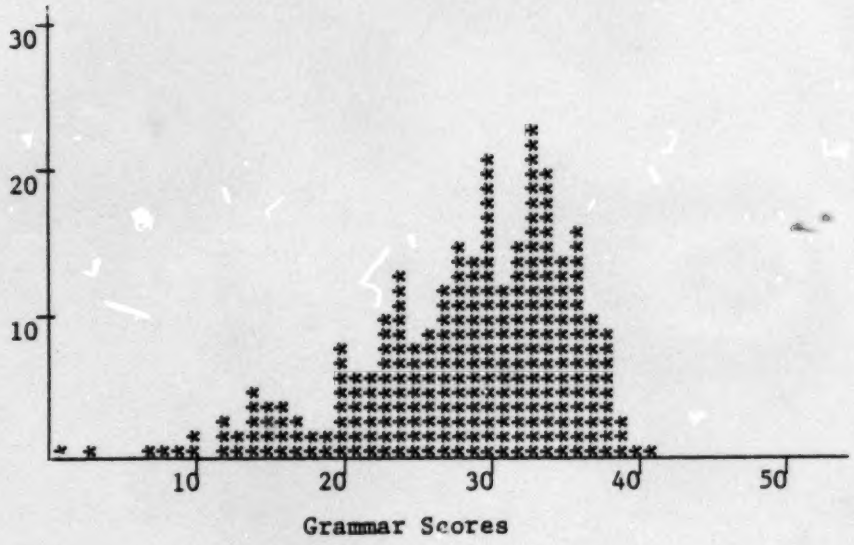

Figure 1._ Distribution of grammar subtest scores, Spring 1981, $\bar{X}$ = 29.31, s.d. = 7.31, N = 278, KR-20 = .88.

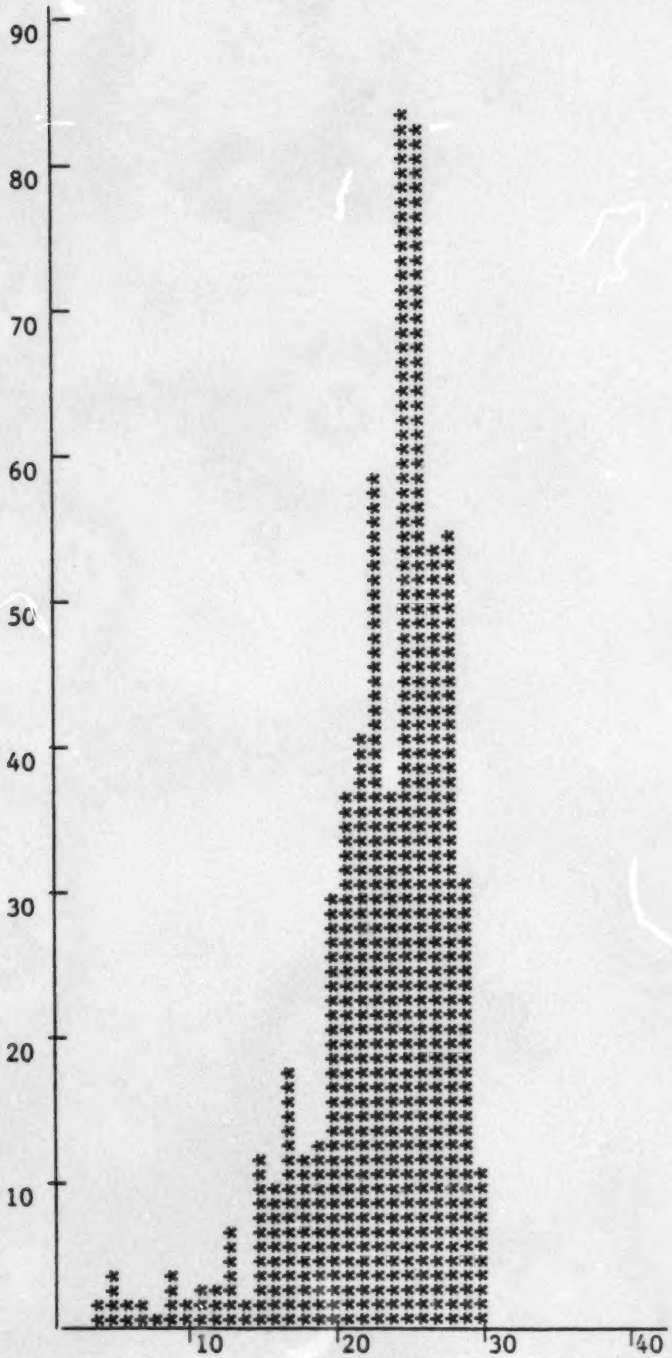

Figure 2. Distribution of old grammar subtest scores, Fall 1981, $\bar{X}$ = 23.31, s.d. = 4.71, N = 629, KR-20 = .824.

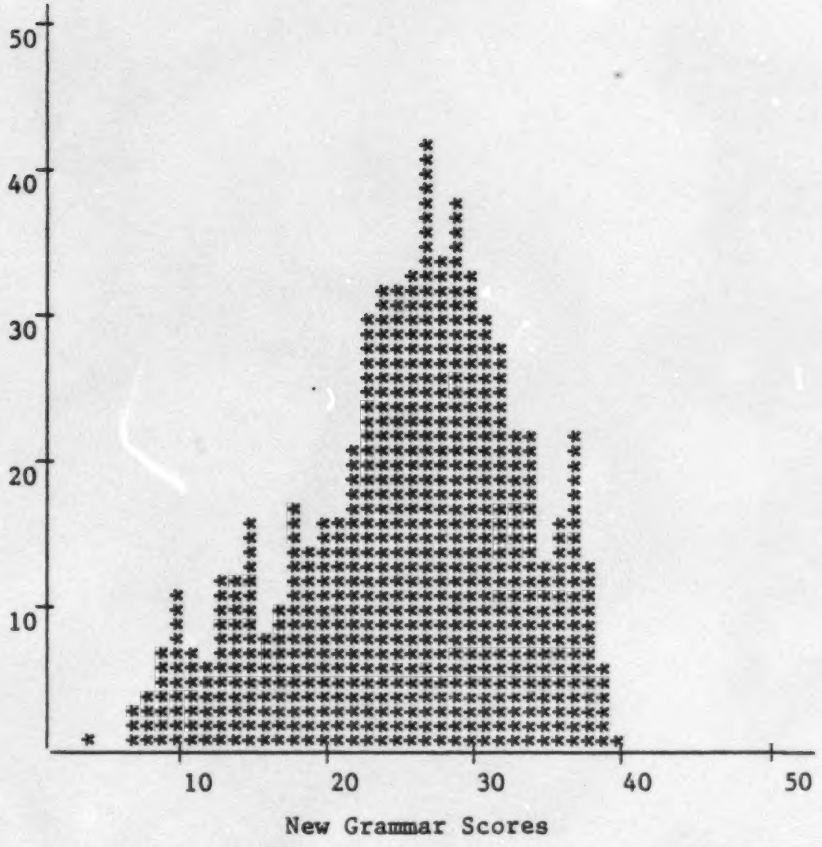

Figure 3. Distribution of new grammar subtest scores, Fall 1981, $\bar{X}$ = 25.70, s.d. 7.46, N = 629, KR-20 = .872.

				Form					
				1			2		
Subtest Area				1	2	3	1	2	3
Items				1. .	. .	. .	. .	. .	. .k
Language Learning Background Group	1	Persons	1						
	2		. . . .						
	3		. . . . . N						

Figure 4. Generalizibility design: Persons(Groups) X Items(Forms X Subtests).

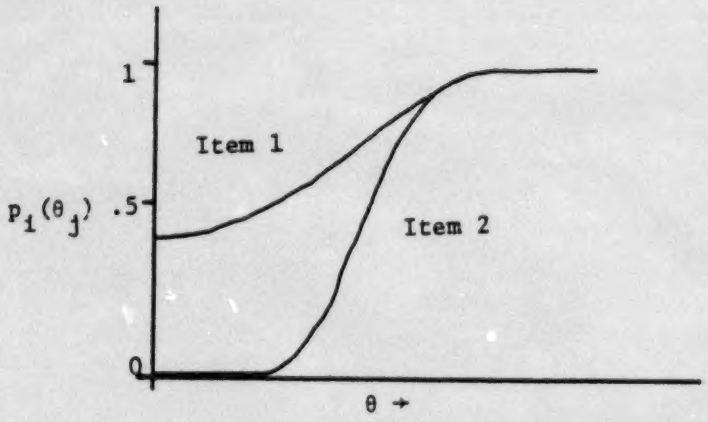

Figure 5. Curves for two different items showing the probability of getting the item correct as a function of θ.

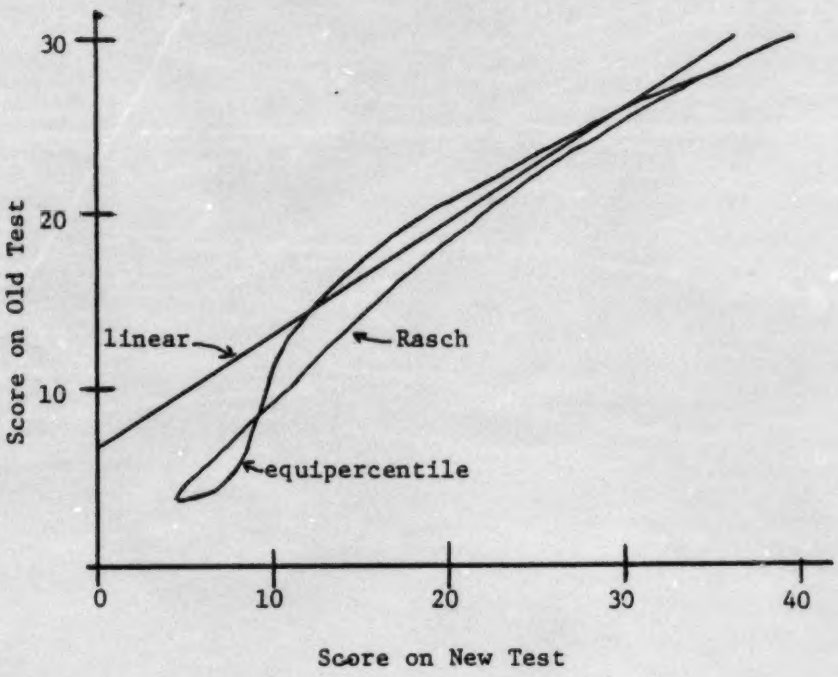

Figure 6. Plot of equated scores derived from linear, equipercentile, and Rasch models.

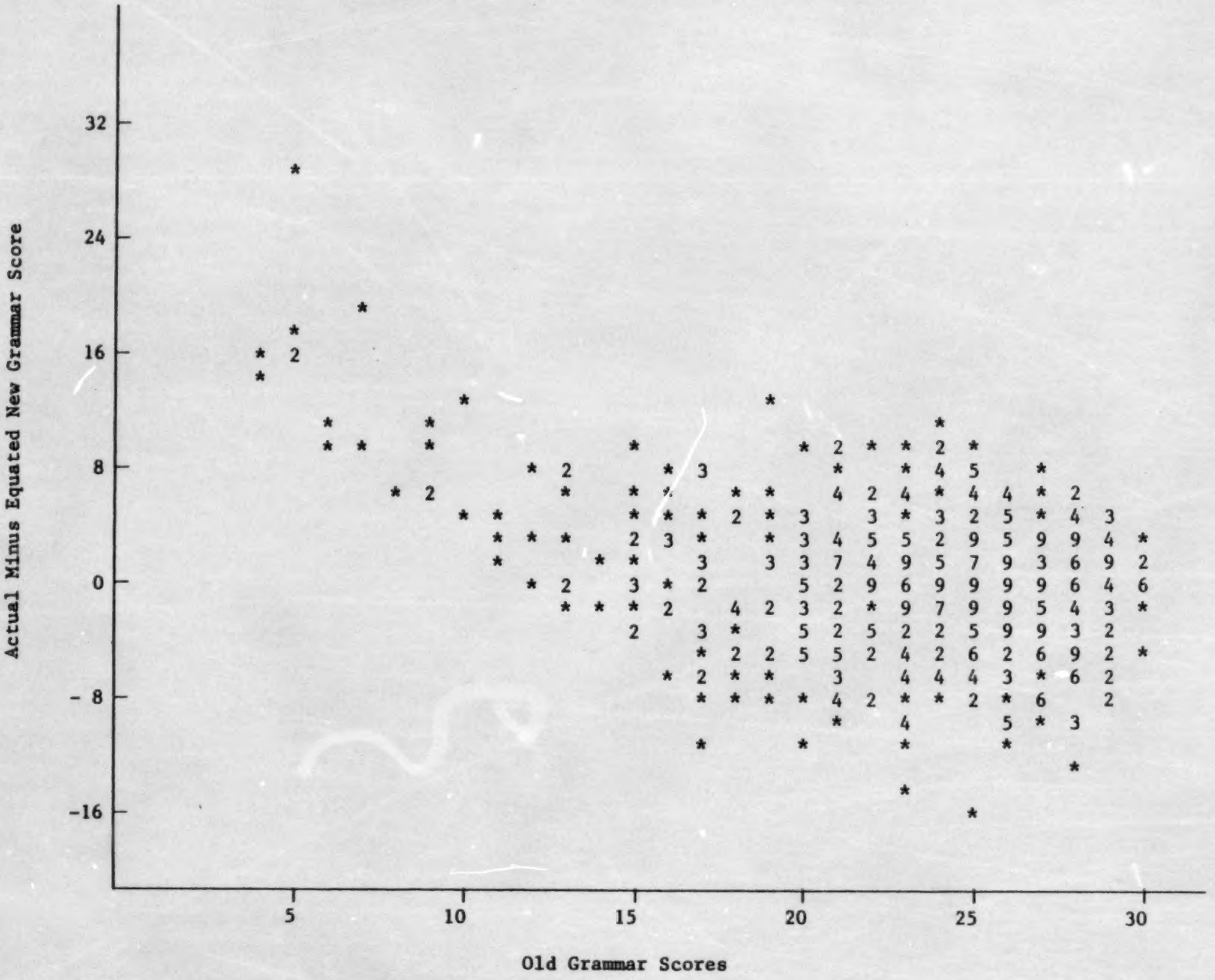

Figure 7. Scatterplot of actual new grammar scores minus equated new grammar scores based on old grammar scores by the linear method of equating.

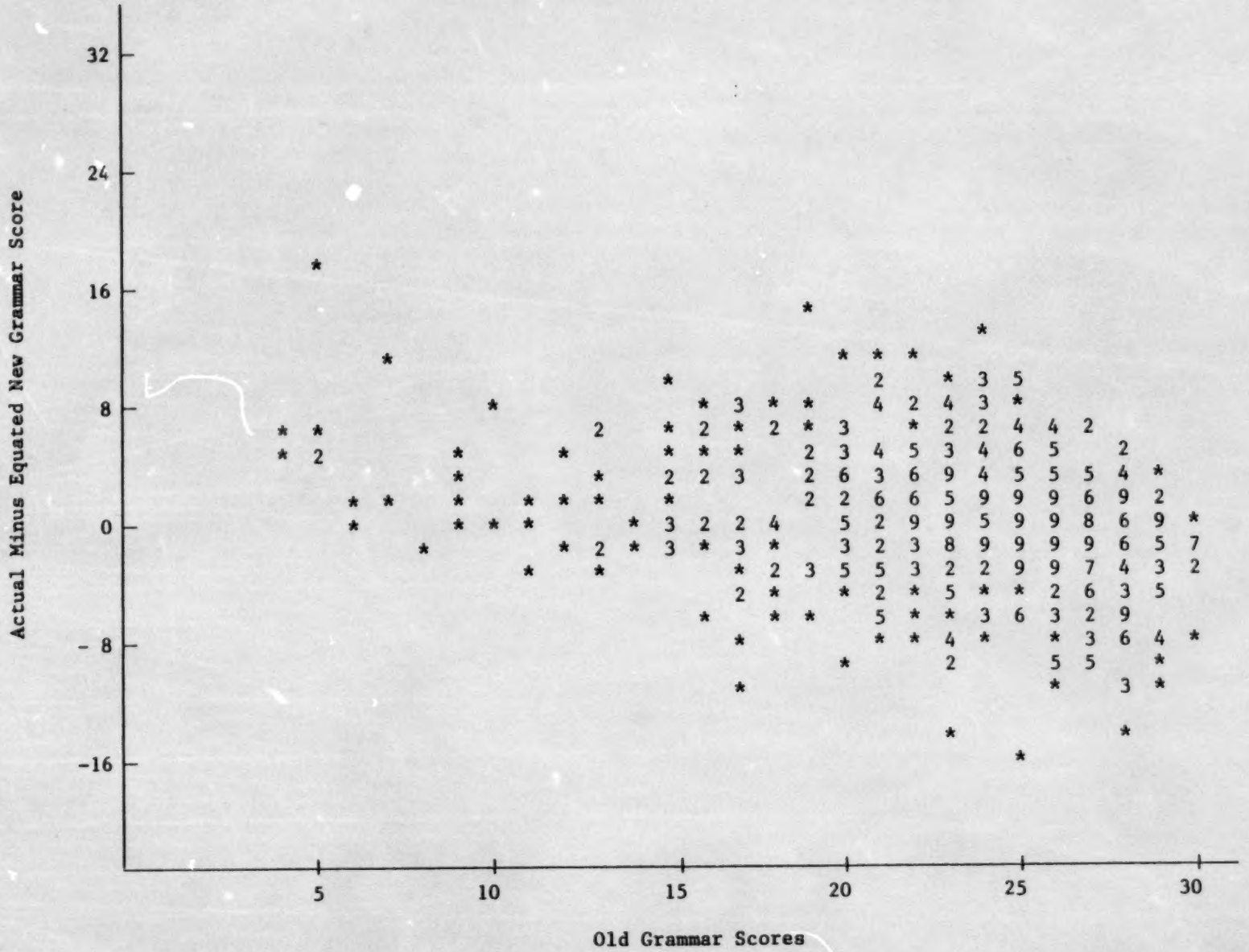

Figure 8. Scatterplot of actual new grammar scores minus equated new grammar scores based on old grammar scores by the equipercentile method of equating.

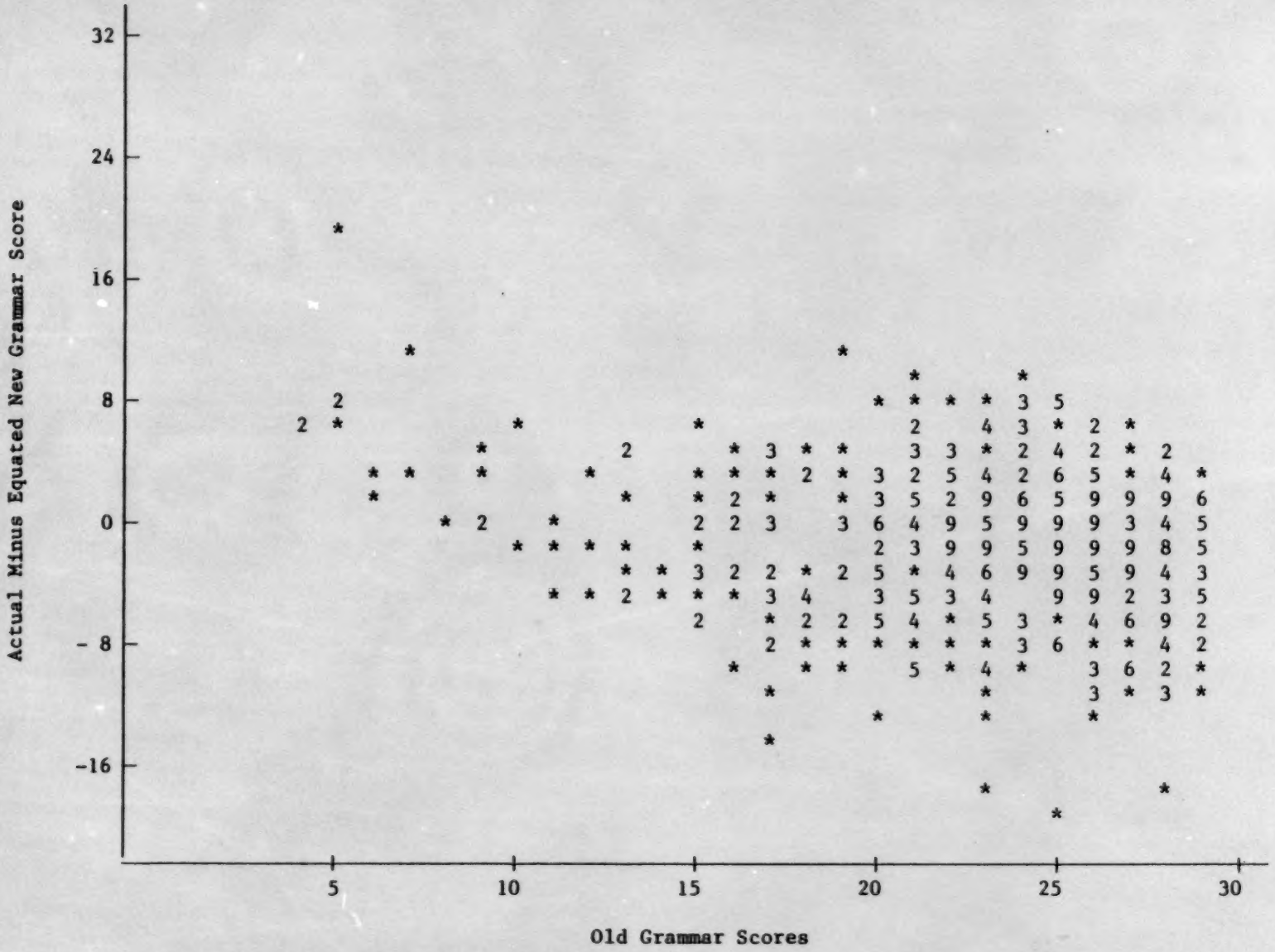

Figure 9. Scatterplot of actual new grammar scores minus equated new grammar scores based on old grammar scores by latent trait equating.

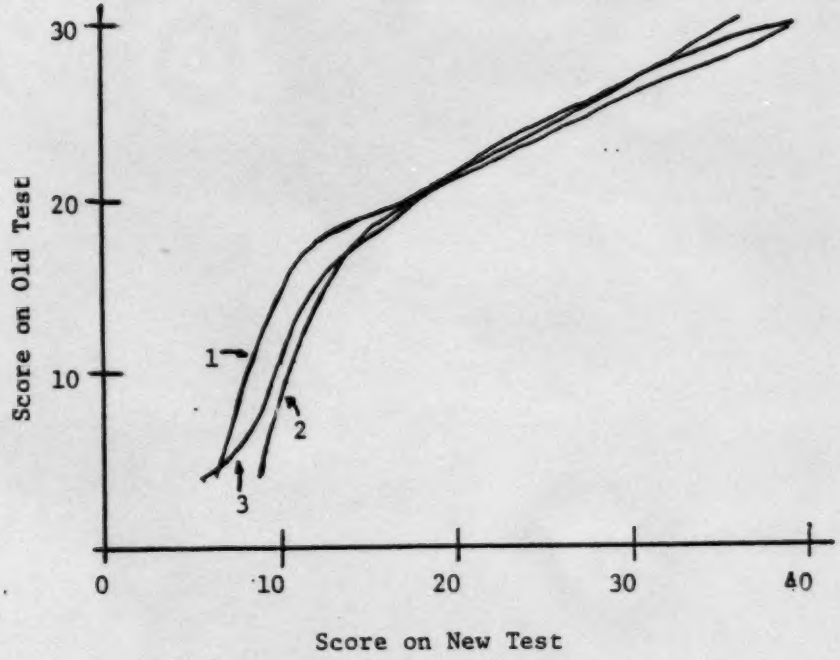

Figure 10. Plot of equipercentile scores derived separately from European academic (1), non-European academic (2), and non-academic (3) ESL students.